Bridging Ontologies and Folksonomies to Leverage Knowledge Sharing on the Social Web: a Brief Survey

Freddy Limpens, Fabien Gandon
Edelweiss, INRIA Sophia-Antipolis, France
{freddy.limpens, fabien.gandon}@sophia.inria.fr

Michel Buffa
KEWI, Laboratoire I3S, Université de Nice, France
buffa@unice.fr

Abstract

Social tagging systems have recently became very popular as a means to classify large sets of resources shared among on-line communities. However, the folksonomies resulting from the use of these systems revealed limitations : tags are ambiguous and their spelling may vary, and folksonomies are difficult to exploit in order to retrieve or exchange information. This article compares the recent attempts to overcome these limitations and to support the use of folksonomies with formal languages and ontologies from the Semantic Web.

1 Introduction

To share and index the large number of resources available on the Web raises several issues that systems based on folksonomies, such as del.icio.us for sharing bookmarks, have recently tried to address. On the other hand, the Semantic Web aims at supporting the exchange of information by developing the interoperability between applications available on the Web. To this end, several methods, tools and principles are proposed, among which formal ontologies play a central role. Generally speaking, ontologies are knowledge representations aiming at "specifying explicitly a conceptualization" (Gruber, 1993). More specifically, formal ontologies use formal semantics to specify this conceptualization and make it understandable by machines. The obstacles to a generalization of ontologies lie mainly in their cost of design and maintenance.

The Web 2.0 consists essentially in a successful evolution of the Web supported by some principles and technologies. Social tagging and the resulting folksonomies (Vanderwal, 2004) are one of the technologies which leveraged Web 2.0 applications. The simplicity of tagging combined with the culture of exchange allows the mass of users to share their annotations on the mass of resources. However, the exploitation of folksonomies raises several issues (Mathes, 2004; Passant, 2007) : (1) the ambiguity of tags, for one tag may refer to several concepts ; (2) the variability of the spelling, for several tags may refer to the same concept; (3) the lack of explicit representations of the knowledge contained in folksonomies; (4) the difficulties to deal with tags from different languages.

Research has been undertaken to tackle the problems posed by the annotation and the exchange of the resources on the Web. The systems or methods they propose strive to reconcile ontology-based models and folksonomy-based models. In section two we present the approaches aimed at extracting the semantics from the folksonomies. In section three we focus on the contributions that support the use of folksonomies with the help of ontologies. In section four, we present some systems exploiting the formalisms of the Semantic Web to assist the exchange of knowledge, and in section five we conclude with a discussion about further investigation in this research topic.

2 Structure in Folksonomies

In this section we focus on the analysis of the semantic information potentially contained within folksonomies, that is what can be inferred from the tags and their usage. In this case, the idea is to keep the simplicity of social tagging interfaces and to infer extra information thanks to finely tunned statistical analysis or data-mining algorithms and additional information from the Semantic Web.

2.1 Building "lightweight ontologies"

Mika (2005) proposes looking at folksonomies as semantic structures emerging from the usages of the com-

munities. He suggests building out of folksonomies "lightweight ontologies" by providing for semantic relationships between the tags. To achieve this task, Mika builds different kinds of networks in order to group related tags. One of these networks allows grouping similar tags by looking at overlapping communities of interest, that is groups of users using the same tag. In this case, a community of interest may be represented by all the actors having used the tag "fishing". If the communities of interest "fishing" and "nautic activities" have a number of actors in common, the tags "fishing" and "nautic activities' will be considered as semantically related. Furthermore, if the group of actors using the tag "fishing" is a subset of the group of actors using "nautic activities", "nautic activities" will be set as a broader term than "fishing".

Lux & Dsinger (2007) have also attempted to extract an ontology out of a folksonomy. Similarly to Mika (2005), they first build a network of tags based on their co-occurrence. Then, they combine a measure of the edit-distance between tags and their co-occurrence to filter wrongly written tags and to merge together similar tags. For example, the tags : "mp3", "audio", and "music", or "game" and "games" are merged together. As a result, the authors obtain a term network which connects together terms extracted from the originating tags. For each term they apply a clustering technique to all the tags co-occurring with the term. The clusters of tags that are obtained are considered by the authors as sub-groups of each term that define each a different context or meaning of that term.

Both methods presented above are deriving semantic relationship between terms out of the tags of a folksonomy, but the relation between the terms are still not defined as precisely as in formal ontologies.

2.2 Dynamic analysis

Halpin *et al.* (2007) analyze the dynamic of folksonomies and look for distribution laws in the frequency of use of the tags. Their hypothesis is that the most used tags to annotate a resource remain the same after a certain amount of time, and their distribution follows a power law. They verify that hypothesis for the seven to ten tags most often associated to popular Web resources posted on a social boomarking service[1]. On the other hand, the authors looked for semantic relationships between the most used tags with the help of inter-tag correlation graphs. Each node of these graphs represent a tag as a circle whose diameter is weighted by the frequence of occurrence of this tag. The length of the edges of these graphs are weighted by their degree of cooccurrence. This visualization is seen as a tool for assisting the construction of ontologies starting from folksonomies.

[1] http ://del.icio.us

2.3 Clustering and mapping with ontologies

The method proposed by Specia & Motta (2007) consists in grouping tags into clusters, and to map these tags to concepts found in ontologies available on the Semantic Web. The clustering is based on how often tags co-occur on the same resource. Then, the system looks for elements from ontologies which have the same label as the clustered tags. In case of success, the system is able to map the concepts and their properties to the tags. The result is a set of clusters of tags enriched with semantics. An attempt to automate this method has been lead by Angeletou *et al.* (2007).

2.4 Data mining and folksonomies

Other works proposed to apply data mining methods to the tripartite model of folksonomies in order to retrieve information in their structure. Jäschke *et al.* (2008) proposed to use formal concept analysis techniques in order to discover the subsets of users sharing the same conceptualizations on the same resources. To do so, they build triples of sets (resources, User, Tags) called tri-concepts where each user has tagged each resource with all the tags. According to the authors, extracting tri-concepts from folksonomies is a first step to buil more structured ontologies from folksonomies. Ontologies are thus seen as social constructions where each concept is described by a set of tags which belong to a set of users and are used to characterize a certain kind of resources. Other data mining techniques has been applied by Schmitz *et al.* (2006) to extract association rules from folksonomies. The first step is to project the tripartite model (resources, Users, Tags) onto a two-dimension structure. For instance, one can consider all the tuples (Users, resources) associated to a set of tags T_x. An example of association rule that may be derived from this projection is : all the users associating tags from the set T_A to a set of resources, oftentimes associate the tags from the set T_B to the same set of resources. This kind of association rule may be exploited in a recommendation system.

3 Enriching folksonomies

In this section we present several works that propose to support folksonomy-based social platforms with the formalisms or the tools of the semantic web. They either use the tags as attributes of the concepts of an ontology (Passant, 2007), or they reify the tags themselves by creating an "ontology of folksonomy" (Gruber, 2005), allowing to get richer metadata from the tagging activity (Tanasescu & Streibel, 2007).

3.1 Guiding tagging with ontologies

Recently, several solutions have been proposed which aim at integrating the least intrusively a tagging interface and knowledge representations based on formal ontologies. Passant (2007) proposes strengthening the social tagging interface of a weblog with a centralized ontology. The idea is to disambiguate while tagging by suggesting to users to connect the terms with which they are tagging to a controlled vocabulary. Thus, if a tag corresponds to two different concepts (for instance the tag "RDF" may correspond to "Resources Description Framework" or to "Reason Distortion Field"), the system asks the user to choose the appropriate concept. When a concept does not exist, users are free to propose a new one to the administrators, who in turn will put it in the right place in the ontology. Social tagging is seen here as en empowerment of the construction of an ontology, which in turn helps disambiguating the possible meanings of a tag.

3.2 Building an ontology of folksonomy

In his article, Gruber (2005) states that there is no opposition between ontologies and folksonomies and proposes constructing an "ontology of folksonomy". The "TagOntology" is a project of an ontology dedicated to formalizing the act of tagging. This model brings in four entities to describe tagging : the tagged object or resource; the term used to tag; the user tagging; and the domain in which the tagging takes place (it can be the service used for instance). Unlike Passant (2007), to whom tags are simple character strings linked to concepts of formal ontologies, Gruber suggests reifying the tagging and to consider each tag as an object as such. To tackle the problems of ambiguity or misuse of tagging (like spam), Gruber proposes to "tag the tags" (as Tanasescu & Streibel (2007) did later, see below). It would then be possible to state that this tag is the synonym of this other tag, or that this tag does not suit this object, integrating mechanisms of regulation like those observed on Wikipedia.

3.3 Getting users to contribute

Tanasescu & Streibel (2007) applied the idea of Gruber (2005) and extended social tagging systems with the possibility to tag the tags themselves and the relationships between them. Indeed, classical tagging system allow their users to add a "tagging relationship", that is a "is_tagged_by" link between a keyword and a document or a web resource. But richer information may be obtained from the tagging activity, like the relationships between the tags. These tagging can easily be expressed with triples, such as "car" - "is_a" - "vehicle", all these tags being freely added by the users. This feature allows exploiting the technologies of the Semantic Web to assist navigation and to suggest to the user other terms semantically related to her query. To prevent irrelevant contributions, the authors proposed solutions based on votes for some tags, in order to appreciate or depreciate them, or solutions based on points that will be granted either to contributors to the tagging task, or to evaluators of the tags of others. Other incentives to contribution could be provided with activities presented as games, but exploited for a utilitarian purpose, such as labeling a huge amount of images on the Web (espgame.org).

4 Ontologies and knowledge sharing

Web 2.0 culture and the technologies of the Semantic Web both focus, however differently, on sharing meaningful information. The complementarity of these approaches, underlined by Greaves & Mika (2008), can be illustrated with the approaches we present in this section. These contributions aim at interconnecting web sites used by on-line communities by exploiting the formalisms of the Semantic Web. The ontologies used in these systems do not describe a particular field of knowledge, but rather the structure of the communities and the interactions between their members and the content they exchange.

4.1 Interlinked on-line communities

The Semantically Interlinked On-line Communities (SIOC) project (Breslin *et al.*, 2005) provide developers social web platforms a formal and technological framework to describe the resources exchanged within and across on-line communities. The formal scheme they propose uses other ontologies like the Simple Knowledge Organization Scheme (SKOS, w3.org/2004/02/skos/) which describes systems of organization of knowledge, and Friend Of A Friend (FOAF, foaf-project.org/) (Brickley & Miller, 2004) which describes the multiple identities and acquaintances of a user. SIOC describes the most common elements present on web sites of communities: the concept of "site", the concept of "post" of a weblog, the concept of "forum", etc. Starting from this vocabulary, the SIOC project proposes tools to automatically annotate the content of some common web applications (e.g. wordpress.org) according to the SIOC ontology.

4.2 Sharing on the Web

Other works propose integrating several ontologies to assist the sharing of data. Hausenblas & Rehatschek (2007) designed "mle", a system which automatically treats mailing lists in order to map the structure of email to appropriate

concepts of an ontology (SIOC). These annotations, generated in RDF, allow this database to be queried with the language of the Semantic Web SPARQL (www.w3.org/TR/rdf-sparql-query/).

Revyu.com (Heath & Motta, 2007) proposes applying the principles of the "Web of Data" to organize the sharing of reviews of cultural items (books, movies, etc.). The "Web of Data" consists in a vision of the Web where the sources of data are located with URIs and interconnected in a decentralized way. Revyu.com includes these principles by (1) allowing anyone to access data stored on other databases in order to prevent redundancies; (2) utilizing RDF to annotate the resources; and (3) keeping open the field of knowledge which can be covered since Revyu.com uses multiple ontologies and other types of knowledge bases to categorize items.

Other approaches allow to semantically structure tagged content to enrich social bookmarking services like GroupeMe![2] (Abel *et al.*, 2007) or inter.est[3] (Kim *et al.*, 2007). "inter.est" uses the SCOT ontology which describe the structure of the tag clouds of the users. The goal of inter.est is to help users find groups sharing the same interests by allowing users to aggregate tag clouds, to form groups of exchange and to facilitate the search of similar tag clouds.

4.3 Semantic Wikis

Semantic wikis were among the first applications to exploit the potential of ontologies to support social tagging practices. SweetWiki (Buffa *et al.*, 2008) is one example: users can edit and modify pages, and also tag any document published on the wiki. The tags are tied together in a folksonomy which is expressed with the languages of the Semantic Web. All the new tags are collected as the labels of new classes which are, by default, subsumed by the class "new concept". All the users are then able to organize the tags of the folksonomy, and to edit them, to add new labels in other languages, to create relations of synonyms, to merge classes, etc.

The author of pages can also use tags to keep an eye on the activity of other contributors in a targeted manner: each user can specify in her homepage her topic of interest in the form of tags. For instance, a user interested in wikis will put a tag "wiki" in the field "interested by". Then, whenever a page is tagged with "wiki" or a term subclass of "wiki", the user will be notified. This function allows watching content that does not yet exist. By keeping track of created or modified pages, and by analyzing over time the behavior of users, it is possible to detect acquaintance networks or communities of interest. This reveals several possibilities: finding the most active person on a given topic, finding the users using similar tags as others, inferring relationships between tags when they are used by the same users, etc.

[2] http://groupme.org/

[3] http://int.ere.st/

5 Discussion

5.1 The best of both worlds

We have seen that it is possible to describe a folksonomy and all the activities occurring on social web sites with an ontology. In this article we have compared different approaches which aim at bridging ontologies and folksonomies to support and leverage the exchange of knowledge over the social web. In that regard, these research works are relevant to the design of social web platforms (which are primarily softwares) in that their methods or algorithms can greatly benefit to the final user's experience, by proposing more precise tools to navigate within and across the different platforms. Interoperability is a critical factor for the future of on-line social softwares, and once adapted to fit the usages, technologies and standards of the semantic web can greatly improve the current situation.

The approaches we presented above often complement each other and they can be distinguished against different criteria which could help describe knowledge exchange systems:

Analysis versus formalization: First, we can extract out of the folksonomies a "lightweight ontology" thanks to statistical analysis (Mika, 2005; Specia & Motta, 2007), or we can directly formalize the tags and their usage among communities of users (Gruber, 2005). Both approaches aim at improving information retrieval in folksonomy-based systems (section 3 and 4).

Type of resource: Second, we can distinguish the different types of resources annotated. Breslin *et al.* (2005) seek to assist the exchange of resources on weblogs and forums, while Heath & Motta (2007) treats the case of reviews. In the same trend, Buffa *et al.* (2008) enhanced the collaborative edition of wiki pages with social tagging functionalities and formalisms of the Semantic Web.

Social context: Third, we can distinguish different kinds of social contexts. A centralized system works well with a clearly defined field of knowledge (Passant, 2007), while, for instance, the collection of reviews of cultural items or bookmarks will require an open field of knowledge (Heath & Motta, 2007).

Design aspects: Fourth, we can distinguish the systems with respect to the design aspects. Some approaches can seamlessly integrate current social platforms such as the SIOC plug-ins, which generate metadata about

the content organized by some popular Content Management Systems (Wordpress, Drupal). Other works can also simply exploit the data already available (Mika, 2005; Halpin *et al.*, 2007) and infer extra semantic information which can in turn be used to describe more precisely the users' data. Finally, other works propose reconsidering the design of social platforms by embedding in them technologies or formalisms of the semantic web (Abel *et al.*, 2007; Heath & Motta, 2007; Buffa *et al.*, 2008).

5.2 Social aspects

It is also necessary to keep in mind the social aspects of knowledge sharing, and to strive to design models fitting actual usages. For example, Sinha (2006) proposed a social and cognitive analysis of tagging. She mentions the distinction brought by Mathes (2004) between the personal use of social bookmarking services, where tagging is used to further retrieve one's own resources, and the social-oriented usage, where the tag is chosen to describe without ambiguity. Sinha also shows that annotating a resource with several keywords requires less cognitive effort than choosing a unique category. Tagging is thus simpler since it allows picking up all the concepts first activated in the mind.

In cases where there exist some contradictions in the different views on the field of knowledge of a community, Zacklad *et al.* (2007) suggest the use of semi-structured ontologies, called "semiotic ontologies" and written following the "Hypertopic" model. Semiotic ontologies still require some skills in knowledge representation, and so they do not constitute an alternative as spontaneous as folksonomies. Yet, they can be considered as an intermediary representation to formal ontologies, in that they are not extended by a "referential formalization". The originality of this approach is to consider the negotiation processes as the main issue; the goal is not to obtain a formal and operational scheme, but rather topic maps or "description networks" (Cahier *et al.*, 2005).

5.3 Perpectives

Gruber (2008) differentiates collective intelligence from collected intelligence. He gives three characteristics of the current systems which collect knowledge: (1) the production of content performed by the users, (2) a synergy between users and the system, (3) increasing benefit with the size of the domain covered. In order to upgrade this type of system towards a collective intelligence, Gruber proposes adding another feature: the emergence of knowledge beyond the mere collection of each contributor's knowledge. He suggests that this fourth feature directly benefit from the integration of the technologies of the Semantic Web.

Thus, the potential of hybrid systems which exploit the benefit of both the ease of use of folksonomies and the support of the formalisms and the methods of the Semantic Web, opens new perspectives for assisting knowledge exchange on the social Web. But several challenges remain. **?** showed the efficiency of combining statistical techniques with extra knowledge from the ontologies on the semantic web, but since the fields of knowledge that could be appropriate is potentially infinite, we need methods to efficiently select the source of information to help structuring the folksonomies. For instance, Review.com Heath & Motta (2007) uses that kind of technique to clearly identify whether the provided web link is about a movie by querying the IMDB.com database, but identifying concepts dealing with the content of the reviews may be more complex, and poses the problem of the selection of the sources of additional information. These issues, plus the need to find similarities between groups of tags or to match tags with element from other ontologies could also benefit from exploiting some of the "ontology matching" field's methods(Euzenat & Shvaiko, 2007). The other challenge that social on-line platforms may be faced with, is the work load needed to administrate or contribute to the system. To add semantic information to the resources exchanged in the social web, the current approaches are: (1) organizing tag data a posteriori, that is analyzing the tags and their usage(Mika, 2005; Specia & Motta, 2007), or proposing the users to organize the tags (Buffa *et al.*, 2008) or tag the tags themselves(Tanasescu & Streibel, 2007); (2) asking users to raise the ambiguity at tagging time (Passant, 2007), or to provide more detailed information when submitting content (Heath & Motta, 2007). The question social software designers may ask at this moment is how much effort they can expect from their users. And this question is not simple since the social context plays a role: incentives to contribute to an enterprise weblog or to a shared reviews platform may largely differs in the amount of effort users may put in providing precise information (workmates may be rewarded for good quality contributions), and (even more complex) in the agreement they may find when dealing with non-consensual knowledge (when commenting on a movie, different and contradictory views may emerge). One of the key to these questions may rely on a balance between top-down-style administration of the knowledge base and bottom-up-style auto-regulation. But both of these components of social software will need appropriate tools to achieve a compromise between the diversity and the precision of the knowledge representations supporting the activities of the "interconnected on-line communities" of the social web.

References

ABEL F., FRANK M., HENZE N., KRAUSE D., PLAPPERT

D. & SIEHNDEL P. (2007). Groupme! - Where Semantic Web meets Web 2.0. In *ISWC/ASWC*, volume 4825 of *LNCS*, p. 871–878: Springer.

ANGELETOU S., SABOU M., SPECIA L. & MOTTA E. (2007). Bridging the Gap Between Folksonomies and the Semantic Web: an Experience Report. *Proc. of ESWC workshop on Bridging the Gap between Semantic Web and Web*.

BRESLIN J., HARTH A., BOJARS U. & DECKER S. (2005). Towards Semantically-Interlinked Online Communities. In *ESWC 2005*.

BRICKLEY D. & MILLER L. (2004). *FOAF Vocabulary Specification*. Namespace Document 2 Sept 2004, FOAF Project. http://xmlns.com/foaf/0.1/.

BUFFA M., GANDON F., ERETEO G., SANDER P. & FARON C. (2008). SweetWiki: A semantic Wiki. *J. Web Sem.*, **6**(1), 84–97.

CAHIER J.-P., ZAHER L., PÉTARD X., LEBOEUF J.-P. & GUITTARD C. (2005). Experimentation of a Socially Constructed "Topic Map" by the OSS Community. *workshop on Knowledge Management and Organizational Memories, IJCAI-05*.

EUZENAT J. & SHVAIKO P. (2007). *Ontology Matching*. Berlin, Heidelberg: Springer.

GREAVES M. & MIKA P. (2008). Semantic Web and Web 2.0. *J. Web Sem.*, **6**(1), 1–3.

GRUBER T. (1993). A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, **5**(2), 199–220.

GRUBER T. (2005). Ontology of Folksonomy: A Mash-up of Apples and Oranges. In *Conference on Metadata and Semantics Research (MTSR)*.

GRUBER T. (2008). Collective knowledge systems: Where the Social Web meets the Semantic Web. *J. Web Sem.*, **6**(1), 4–13.

HALPIN H., ROBU V. & SHEPHERD H. (2007). The Complex Dynamics of Collaborative Tagging. In *WWW*: ACM Press.

HAUSENBLAS M. & REHATSCHEK H. (2007). mle: Enhancing the Exploration of Mailing List Archives Through Making Semantics Explicit. In *Semantic Web Challenge, ISWC*.

HEATH T. & MOTTA E. (2007). Revyu.com: a Reviewing and Rating Site for the Web of Data. In *ISWC/ASWC*, volume 4825 of *LNCS*, p. 895–902: Springer.

JÄSCHKE R., HOTHO A., SCHMITZ C., GANTER B. & STUMME G. (2008). Discovering Shared Conceptualizations in Folksonomies. *J. Web Sem.*, **6**(1), 38–53.

KIM H.-L., YANG S.-K., SONG S.-J., BRESLIN J. G. & KIM H.-G. (2007). Tag Mediated Society with SCOT Ontology. In *Semantic Web Challenge, ISWC*.

LUX M. & DSINGER G. (2007). From folksonomies to ontologies: Employing wisdom of the crowds to serve learning purposes. *International Journal of Knowledge and Learning (IJKL)*, **3**(4/5), 515–528. ISSN (Online): 1741-1017 ISSN (Print): 1741-1009.

MATHES A. (2004). Folksonomies - Cooperative Classification and Communication Through Shared Metadata. *GSLIS, Univ. Illinois Urbana-Champaign*.

MIKA P. (2005). Ontologies are Us: a Unified Model of Social Networks and Semantics. In *ISWC*, volume 3729 of *LNCS*, p. 522–536: Springer.

PASSANT A. (2007). Using Ontologies to Strengthen Folksonomies and Enrich Information Retrieval in Weblogs. In *WSM*.

SCHMITZ C., HOTHO A., JÄSCHKE R. & STUMME G. (2006). Mining Association Rules in Folksonomies. *Data Science and Classification*, p. 261–270.

SINHA R. (2006). Tagging from Personnal to Social : Observations and Design Principles. In *Tagging Workshop, WWW*.

SPECIA L. & MOTTA E. (2007). Integrating folksonomies with the semantic web. *ESWC*.

TANASESCU V. & STREIBEL O. (2007). ExtremeTagging: Emergent Semantics through the Tagging of Tags. In *ESOE at ISWC*.

VANDERWAL T. (2004). Folksonomy Coinage and Definition. http://www.vanderwal.net/folksonomy.html.

ZACKLAD M., BNEL A., CAHIER J., ZAHER L., LEJEUNE C. & ZHOU C. (2007). Hypertopic : une Métasémiotique et un Protocole pour le Web Socio-Sémantique. In *IC*, p. 217–228.